جامعة نيويورك أبوظبي

NYU | ABU DHABI

PSYCH-UH 1004Q: Statistics for Psychology

Class 9: The logic of null hypothesis testing (part 2)

Prof. Jon Sprouse
Psychology

Quick recap

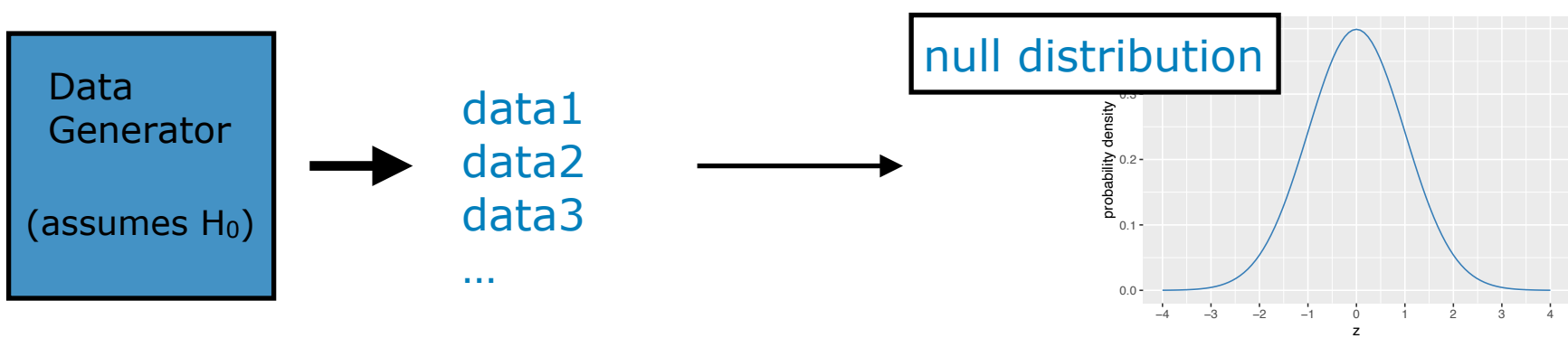
(Fisher’s approach)

The mathematical part of NHT

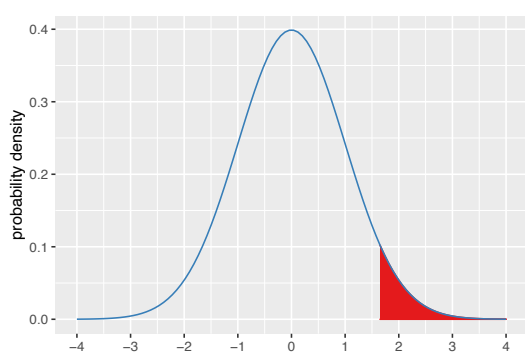
The mathematical part of NHT has three steps:

1. Run an experiment to collect the observed data. Calculate a statistic from it, like the mean or a z-score.

2. Assume that the null hypothesis is true, and generate **all possible data sets** that could arise (using the same sample size as your experiment). We summarize it as a distribution called the null distribution.



3. Look up the probability of the observed data or data more extreme in the null distribution. This is a conditional probability.

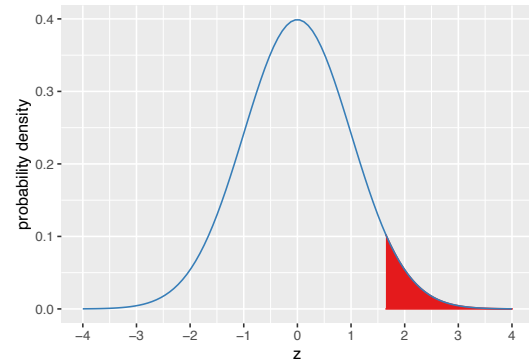


$$\mathbf{P(data \mid H_0)} = \frac{\text{observed data}}{\text{generated data}}$$

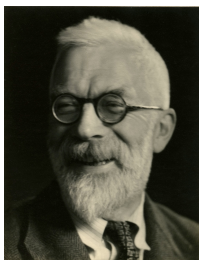
The logical part of NHT

The **mathematical** part of NHT yields a conditional probability - the probability of obtaining the observed data or data more extreme under the assumption that the null hypothesis is true. We call this a ***p*-value**.

$$p\text{-value} = P(\text{data} \mid H_0) = \frac{\text{observed data}}{\text{generated data}}$$



The **logical** part of NHT interprets the *p*-value.



Interpreting *p*-values is actually a fairly philosophical act. We will start with Fisher's philosophy, because he started NHT. His interpretation can be captured in a statement called **Fisher's disjunction** (a disjunction is a statement with "or" in it):

If p(data|H_0), called the p-value, is sufficiently low, then you can conclude either: (i) the null hypothesis is incorrect, or (ii) a rare event occurred.

Making decisions:
The Neyman-Pearson approach to NHT

Two approaches to NHT

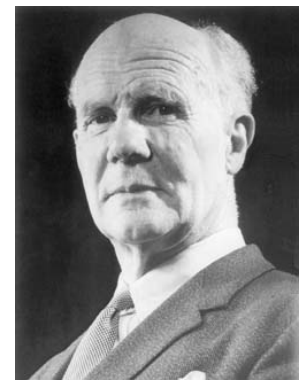
It turns out that there are two major approaches to NHT. They use the same exact math, so it is easy to think that they are identical. But they differ philosophically, so it is important to keep them separated.

Ronald Fisher was the first person to try to wrangle the growing field of statistics into a unified approach to hypothesis testing. His NHT was the first attempt. We have already seen the Fisher approach. For Fisher, *p*-values are a measure of the strength of evidence against the null hypothesis.

Ronald A. Fisher
(1890-1962)

Neyman and Pearson were fans of Fisher's work, but thought he missed an important component - **you must make a decision about whether to reject the null hypothesis or not**. They sought to make the decision process as mathematically concrete as possible.

Jerzy Neyman
(1894-1981)

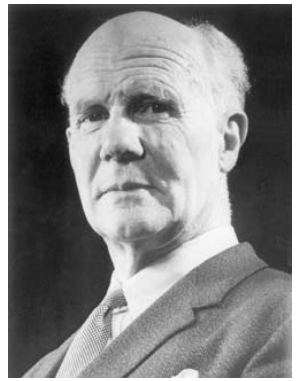


Egon Pearson
(1895-1980)

The Neyman-Pearson approach

Our book adopts the Neyman-Pearson approach. So, now that we have seen the basics with Fisher's approach, I want to lay out the N-P approach, and the concepts that it adds to NHT.

Jerzy Neyman
(1894-1981)



Egon Pearson
(1895-1980)

The fundamental addition is the idea of a **decision**. You must decide whether to reject the null hypothesis based on the *p*-value that you obtain.

This leads to a number of concrete additions to the NHT process for us.

1. A decision criterion called the **alpha level** or **alpha criterion**.

2. A definition of the types of **errors** that can arise: **type I** and **type II**.

3. A theory of the relationship between the **alpha level** and **type I error rate**.

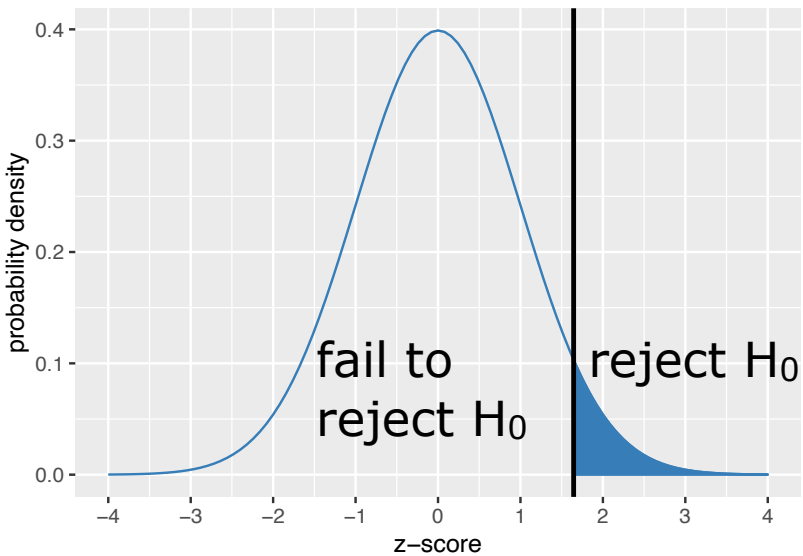
(There are other additions, but we will wait until later in the course to add those.)

The alpha level

(a criterion for your decision)

Setting a criterion for the decision

The first concept that the N-P approach adds to NHT is a decision criterion. The decision criterion is called the **alpha level** or **alpha criterion**. It is typically chosen based on a *p*-value, and then converted to a critical test statistic value.



Here I have set the alpha level to .05 based on the convention in psychology to choose p=.05.

The resulting critical value for the (one-tailed) z-test is 1.645. I have marked it with a vertical line.

If the test statistic for our sample is <u>beyond</u> (moving away from the mean) the critical value determined by the alpha level, we <u>reject</u> H_0.

If the test statistic for our sample is <u>within</u> (closer to the mean) the critical value determined by the alpha level, we <u>fail to reject</u> H_0.

The consequences of your decisions

(a theory of errors)

Decisions can lead to errors

There are two states of the world: the null hypothesis is either true or false.

You can never know if the null hypothesis is true or false. This actually follows from the philosophy of science and the problem of induction.

In the absence of certainty about the state of the world, all you can do is make a decision about how to proceed based on the results of your experiment. You can choose to reject the null hypothesis, or not.

This sets up four possibilities: two states of the world and two decisions.

		the state of the world	
	H_0 is…	True	False
your decision	Rejected	Type I error (false positive)	correct decision (true positive)
	Not Rejected	correct decision (true negative)	Type II error (false negative)

Type I Error: This is when the null hypothesis is true, but you reject it.

Type II Error: This is when the null hypothesis is false, but you fail to reject it.

Type I errors are worse than type II errors

Type I Error: This is when the null hypothesis is true, but you reject it.

Type II Error: This is when the null hypothesis is false, but you fail to reject it.

Though we obviously want to minimize both types of errors if we can, scientists tend to consider type I errors to be riskier.

In a type I error, you conclude that there is something interesting going on (an effect in your experiment) when really there is nothing interesting going on (no effect). In other words, they lead you to postulate a more complex universe.

This is riskier because it wastes everyone's time. If you think there is an effect, you will create a theory for it, and publish it. Others will spend time exploring that theory. It may take time and effort to notice that you were mistaken. Then it will take time and effort to conclusively prove it was a mistake.

Type II errors are less risky. They are just you failing to notice something interesting. That is the default state of the world. You won't write a paper. No one will know. It isn't fun. But it doesn't waste anyone's time!

The type I error rate

Imagine that you are going to run experiments over and over for the rest of your life. One thing you might want to know is the following: out of all of those experiments, how many are type I errors?

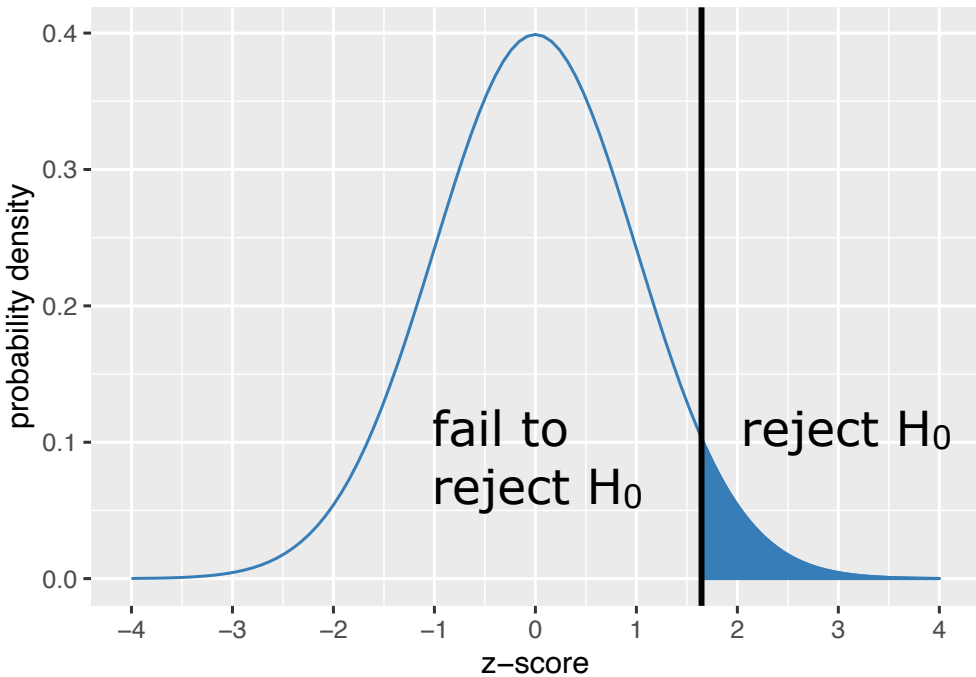
We can call this the type I error rate. If I said my type I error rate is .05, that is me saying that out of all the experiments that I am running, about 5% of them are type I errors.

Neyman and Pearson had the following insight: Though we can never know if any individual result is an error or not (remember philosophy of science, we can never know!), we can try to minimize our long term type I error rate. This will allow us to say, for example, that we only expect 5% of our results to be errors over the long term.

And they demonstrated that the way to do this is to select a decision criterion (the alpha criterion) and apply it consistently over the long term. If you select it appropriately (based on some math), you can control your type I error rate!

The type I error rate is determined by the alpha level that you choose

Here is the critical fact: For a single experiment with a single statistical test, if you choose an alpha level of X, then the type I error rate will be X.



Here I have chosen an alpha of .05. So, my decision criterion is p=.05.

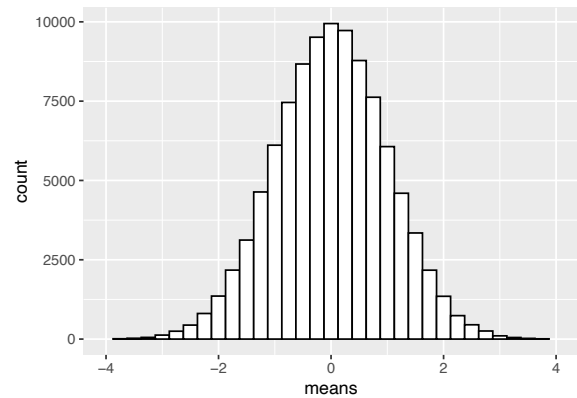
Because this is a single experiment with a single statistical test, this means that my type I error rate will be .05

Warning: Alpha always determines the type I error rate. But it is only equal to the type I error rate for individual tests. For situations where we do multiple tests at once, the relationship is more complicated (but still determined by a regular relationship between alpha and the type I error rate). We will see this later in the course in a section called "multiple comparisons".

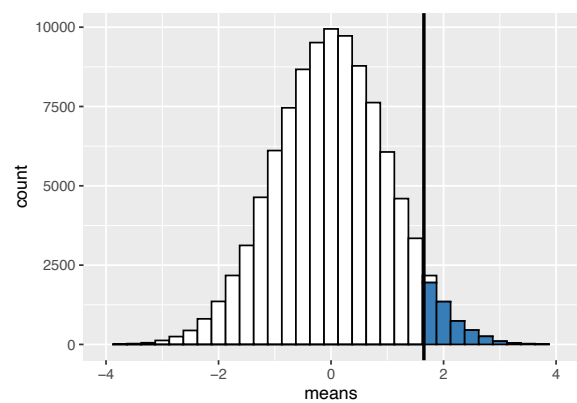
Demonstration: alpha is the type I error rate

Let's simulate a very large number of experiments, as if we were repeating them over and over, and see that the type I error rate is the same as our chosen alpha level.

First, let's simulate 100,000 experiments in which the null hypothesis is true. I'll draw the results as a distribution. I'll use z-scores to keep things simple! Remember, this is as if we ran 100,000 experiments in our life!



Next, let's identify all of the experiments in which a type I error occurs. Since we assumed that the null hypothesis is true, an error will occur in any experiment with a *p*-value less than .05. Since we used z-scores, this will be any z-score beyond 1.645.



Next, we can count the number of experiments in our lives with z values beyond 1.645. Notice that it is roughly .05! (It is a simulation, so it is not precise.)

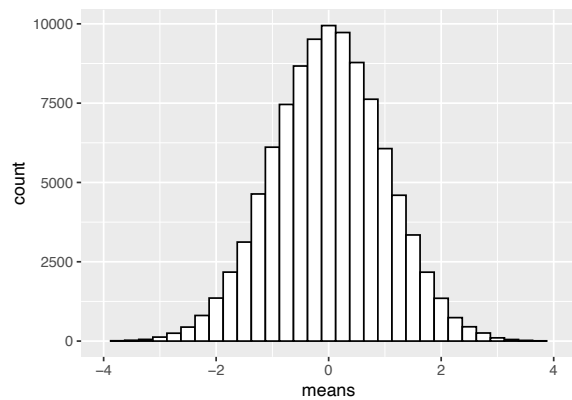
$$\frac{4{,}931}{100{,}000} = .049$$

Why is alpha equal to the type I error rate?

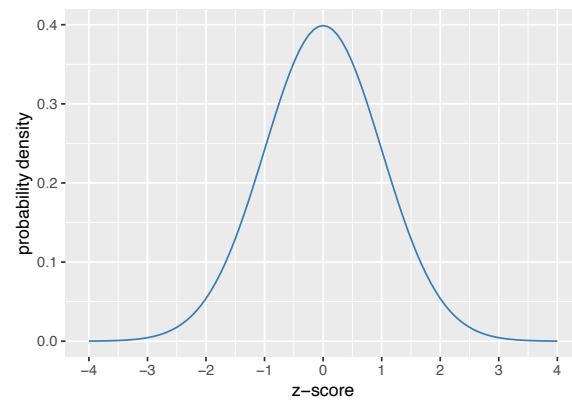
The critical insight for understanding **why** this happens is that the large set of experiments that we simulated all assume that the null hypothesis is true.

This means that our distribution of lifetime experiments is actually equivalent to a null distribution!

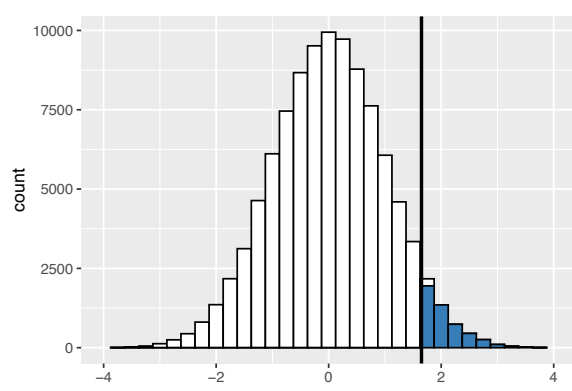
lifetime experiments

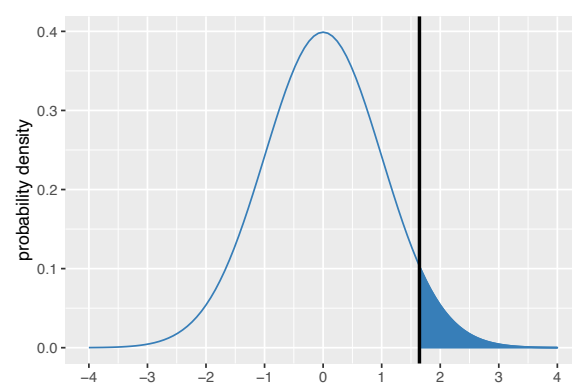


null distribution



So, when we select an alpha level of X (say, .05), we are also selecting the threshold in our distribution of lifetime experiments for determining type I errors.



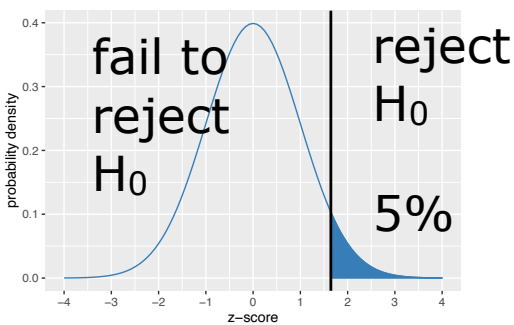


The direction of the alternative hypothesis and the consequences for the probabilities

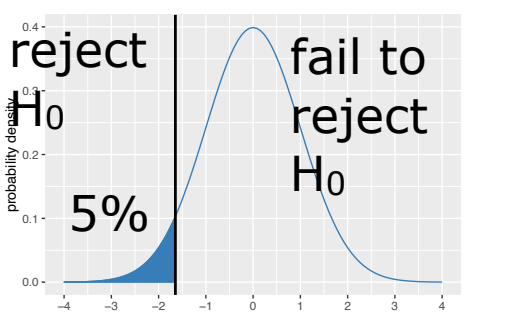
(i.e., one-tailed versus two-tailed tests)

The direction of the alternative hypothesis

Though we do not study the alternative hypothesis directly, it does have some impacts on our methods. For example, if your alternative hypothesis has a specific direction, that will determine which tail of the null distribution counts as the critical region for your test. These are called **one-tailed tests**.



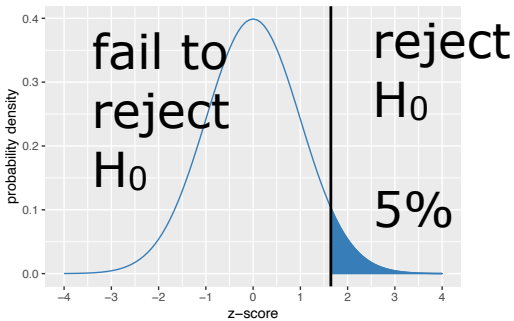
If your alternative hypothesis states that the mean of the sample will be larger than expected under the null hypothesis, the critical tail will be to the right. For z-scores, this means the critical z will be positive!



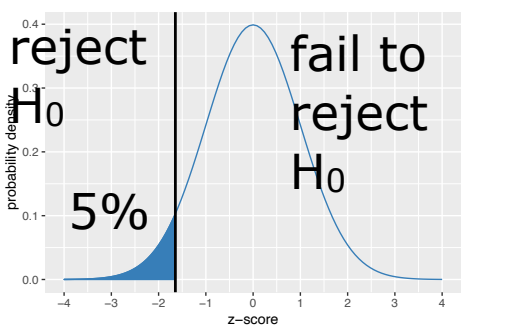
Of course, it is also possible to have an alternative hypothesis that the mean of the sample will be smaller than expected under the null hypothesis. In this case, the critical tail will be to the left. For z-scores, this means the critical z will be negative!

If you have a directional alternative hypothesis, you **must** respect the logic of it. You must only reject the null hypothesis with scores in the selected tail. If you get a score in the opposite tail, you must fail to reject the null hypothesis.

Examples of directional hypotheses, and therefore one-tailed tests



H_1 is that the mean of the sample will be **higher** than expected. Our IQ test study is a great example. Our H_1 is that practicing the test will increase scores. So our sample z will be positive! So we look at the right tail of the distribution, and set a critical value to a z of 1.645



H_1 is that the mean of the sample will be **lower** than expected. Our vaccine study is a great example. Our H_1 is that taking the vaccine will decrease scores. So our sample z will be negative! So we look to the left tail of the distribution, and set a critical value to a z of -1.645.

If you have a directional alternative hypothesis, you **must** respect the logic of it. You must only reject the null hypothesis with scores in the selected tail. If you get a score in the opposite tail, you must fail to reject the null hypothesis.

A non-directional alternative hypothesis

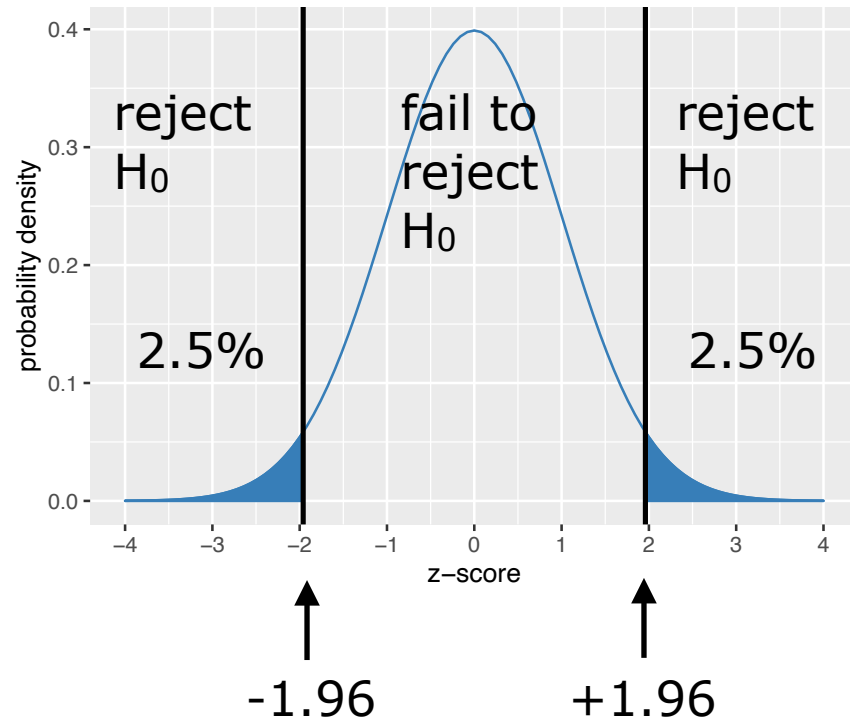
It is also possible to specify a non-directional alternative hypothesis. This would say that the mean of the sample will be either larger or smaller than what is expected under the null hypothesis. These are called **two-tailed tests.**

There are some things to notice about this:

First, this means that there are two critical regions. We have two critical values for dividing the distribution.

Second, we split the critical region between the two tails. If our desired alpha is .05 (therefore a type I error rate of .05), we place .025 in each tail.

Third, this split changes the critical values. The two-tailed critical values for p=.05 are -1.96 and +1.96.



In general, the critical values for a **two-tailed test will always be larger in magnitude** (higher number, regardless of sign) than the critical values for a one-tailed test.

Why do we split the probability between tails?

The splitting of the probability between tails can be a little confusing. But it follows directly from the definition of a *p*-value and the rules of probability!

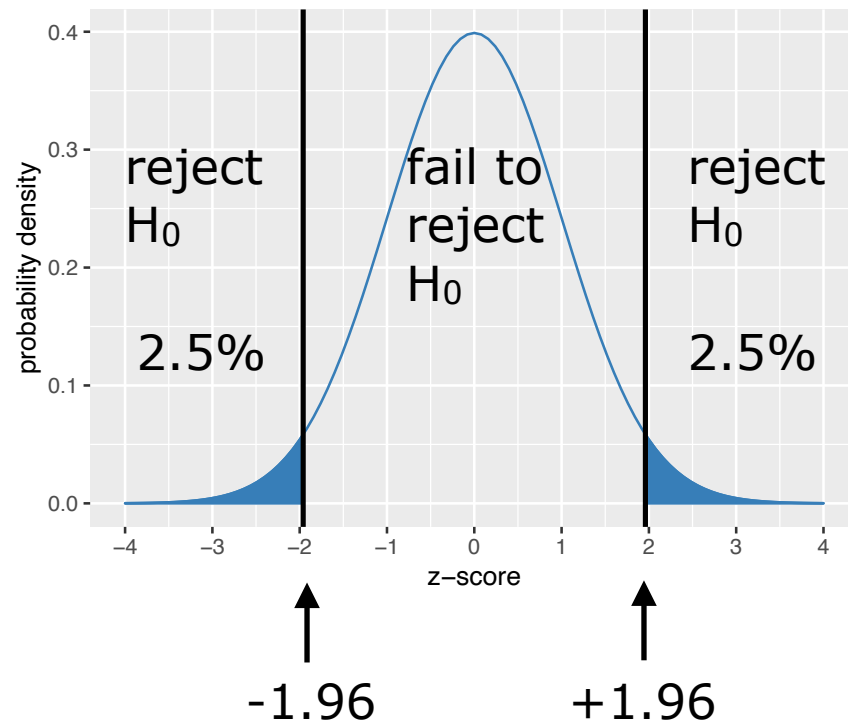
***p*-value:** The probability of obtaining the observed value or a value that is more extreme.

With a non-directional hypothesis, more extreme means either above the mean OR below the mean.

P(A or B): P(A) + P(B)

The "or" rule of probability tells us that we must add together the probability of the two events.

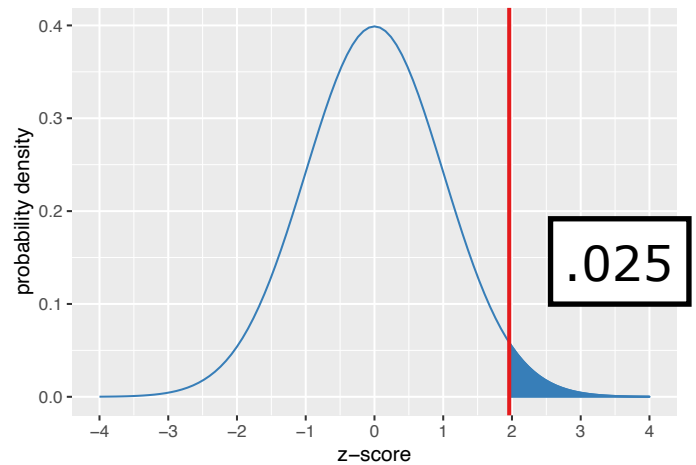
Therefore, our *p*-value will be the sum of the probability in the two tails. If we want P(A or B) to be .05, that means choosing values that put .025 in P(A) and .025 in P(B)! So, we put .025 in each tail!



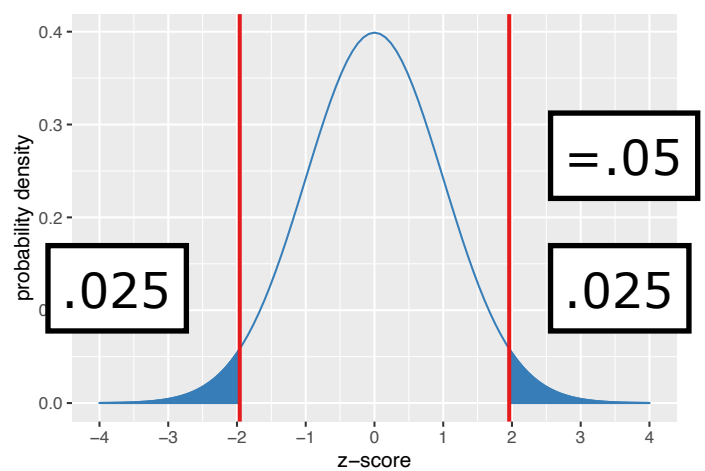
This also means that the *p*-values we calculate from a given score are twice as large

A related consequence of this is that the *p*-values for a two-tailed test will be twice as large as the *p*-values for a one-tailed test with the same score!

Let's say we obtained a z-score of +1.96. With a **one-tailed test**, the probability of obtaining that score or one more extreme is .025. We can see this in Table A1 in our book or by using pnorm() in R.

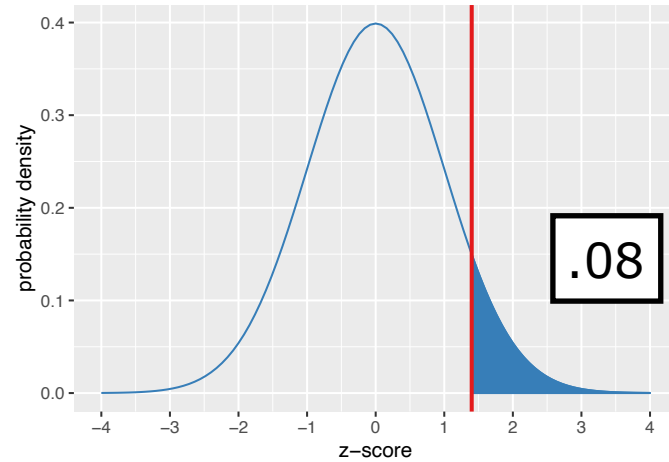


But with a **two-tailed test**, the probability of obtaining that score or one more extreme is .05. This is because we must add the probability from both tails because "one more extreme" also means a z-score less than -1.96.

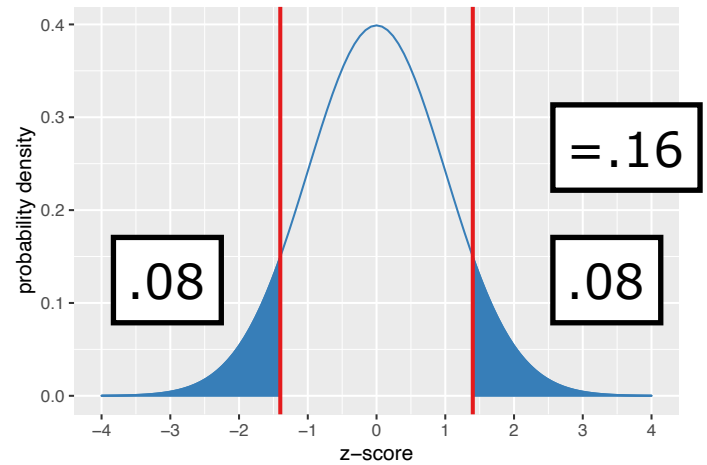


One more example to drive it home

Let's say we obtained a z-score of +1.4. With a **one-tailed test**, the probability of obtaining that score or one more extreme is .08. We can see this in Table A1 in our book or by using pnorm() in R.



But with a **two-tailed test**, the probability of obtaining that score or one more extreme is .16. This is because we must add the probability from both tails, which means adding the probability of a z-score beyond -1.4.



How to **run** a statistical test

(The six concrete steps to running a test in the Neyman-Pearson approach.)

How to **run** a statistical test (N-P approach)

1. State the **null hypothesis** and whether the alternative hypothesis is **one-tailed or two-tailed**.

2. Select the **statistical test** and the **alpha level**.

3. Select the **sample size** and the **collect the data**.

4. Identify the **critical values** for the decision to reject the null hypothesis. (Remember that this depends on one-tailed versus two-tailed!)

5. Calculate the **test statistic** for the observed data.

6. Make the **statistical decision** based on the test statistic and the critical values.

A really big question for NHT:
Why focus on the null hypothesis instead of the alternative hypothesis?

Falsification limits us

The first thing to note is that we are working with falsification, not confirmation. So all we can do is reject one hypothesis at a time.

The second thing to note is that there are actually an infinite number of alternative hypotheses. For example, if you are measuring how many days a COVID vaccine lasts, possible alternative hypotheses are 1 day, 2 days, 3 days… 1 month, 2 months, etc…

The third thing to note is that the null hypothesis is always the most likely hypothesis. Again, think about how many chemical compounds there are in the world. Millions? Billions? Only a few work as COVID vaccines. So for most compounds that one could test, the null hypothesis (0 days) would be true.

Given all of these, let's think about what would happen if we focused on one alternative hypothesis for a new COVID vaccine. Let's say that it protects for 90 days. It would most likely be falsified. But there would still be an infinite number of other alternatives to test (91 days, 92 days, etc). And, we still wouldn't even know if it works at all — most likely it does not (0 days) because most chemical compounds are not effective vaccines.

For all of these reasons the null hypothesis is the most strategic first choice. We need to show that there is something interesting to study!

Math limits us too

Remember that in order to calculate a *p*-value, we need to assume that the hypothesis that we want to falsify is true, and then either simulate or calculate a distribution of hypothetical results.

It is relatively easy to either simulate or calculate a distribution for the null hypothesis. The reason for this is that there is nothing particularly interesting going on. There is no effect. There is nothing. The variation that we see across samples is due to random variation. There is no specific data generation process going on to cause the variation.

But things are not so easy under an alternative hypothesis. If an alternative hypothesis is true, there is a specific process going on. We would need to understand that process in some amount of detail in order to generate the possible results. We can't just assume that the distribution would be normal.

And, we'd have to do this for every new topic of study, and for every new experiment, because the specific processes will change from topic to topic. We couldn't just have one set of distributions like we do for the null hypothesis.

This is not impossible to do. Modern Bayesian methods do something like this. But they rely on complicated assumptions and sophisticated computer simulations. NHT tries to do more with less!

A final important concept for NHT: What a *p*-value is and what a *p*-value is not

False beliefs about *p*-values

A *p*-value is the probability of obtaining the observed data or data more extreme under the assumption that the null hypothesis is true. It is one specific piece of information.

$p(\text{data} \mid H_0)$

There are lots of other pieces of information about an experiment or hypothesis that scientists often want to know. And they are often probabilities.

1. The probability of the null hypothesis being true: $p(H_0 \mid \text{data})$

2. The probability of your hypothesis of interest being true: $p(H_1 \mid \text{data})$

3. The probability of incorrectly rejecting the null hypothesis: $p(\text{sig.} \mid H_0)$

4. The probability that you can replicate your results with a second experiment: $p(\text{data2} \mid \text{data1})$.

It is critical to realize that *p*-values are not any of these. Sometimes people falsely believe that they are, but that leads to errors in reasoning. The best way to avoid this pitfall is to memorize what a *p*-value is (the equation at the top). Then, whenever you find yourself wanting a different probability, write out both the *p*-value equation and the equation for the new probability, and see if they are the same!